POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Detection of Illegal Landfills using Deep Learning: A Weakly Supervised Approach

Laurea Magistrale in Computer Science and Engineering - Ingegneria Informatica

Author: Simone Sorrenti

Advisor: Prof. Giacomo Boracchi

Co-advisor: Prof. Piero Fraternali, Ph.D. Andrea Diecidue

Academic year: 2023-2024

1. Introduction

In the contemporary era, the detection of illegal landfills presents a critical challenge, with hidden sites posing significant environmental and public health risks. Through the utilization of satellite imagery, this research aims to identify and locate illegal landfills using binary classification and segmentation techniques. Aligned with the European PERIVALLON project [2, 6], this study contributes to fight organized environmental crime with cutting-edge artificial intelligence solutions.

While traditional on-site inspections remain crucial, monitoring large areas presents significant challenges. Recent studies have shown that advanced aerial imagery analysis technologies offer a promising approach for swiftly identifying potential illegal landfills. Additionally, Computer Vision techniques, particularly Deep Learning, facilitate the development of automated tools for recognizing subtle landfill characteristics, enhancing detection efficiency [3].

However, the scarcity of semantic segmentation annotations poses a significant hurdle in training advanced deep learning models. To address this limitation, approaches like Weak-Self Supervision and Data Augmentation techniques have emerged, offering viable alternatives for training models using only label-level annotations.

Other challenges with satellite imagery are the presence of small objects dispersed throughout complex backgrounds, varying sizes, orientations, and high intra-class diversity.

My research focuses on optimizing image preprocessing techniques, implementing a data augmentation framework, and exploiting CAM and MIL techniques for weakly supervised localization. Specifically, we standardized the spatial resolutions and dimensions of images from the AerialWaste dataset [8] to facilitate consistent object identification. We developed a data augmentation framework to simulate satellite conditions and enhance model performance. Lastly, we compared the effectiveness of CAM and MIL techniques in identifying regions of interest in weakly supervised learning scenario.

CAM relies on the concept of using class activation maps produced by a convolutional neural network to identify regions of interest in the image. Regions with higher activations are considered as regions of interest for that class. On the other hand, MIL is based on learning from sets of extracted patches rather than entire image. During training, the model seeks to learn which sets of patches are associated with a particular class, without knowing the specific label for each patch.

2. Problem formulation

The problem of semantic segmentation in waste detection can be formally described as follows: Given an image $I \in R^{W \times H \times C}$ and a set of classes $C = \{c_1, c_2, .., c_k\}$, find a function $g : R^{W \times H \times D} \rightarrow C^{W \times H}$ that associates to each pixel a unique class $c_i \in C$. The function g produces a mask $\Delta \in C^{W \times H}$ in which for each pixel we have a unique value that represents the class to which it belongs.

To adress this problem, we utilized the AerialWaste dataset [8], which comprises satellite images of the Lombardia region provided by the ARPA agency in Italy. In its version 2.0, the dataset includes a total of 10,977 RGB satellite images, meticulously divided into training and testing sets to facilitate precise comparisons between various studies. Among these images, there are 7,318 negative samples and 3,659 positive samples.

The images available do not all come from the same source; rather, we can identify three different sources: AGEA Ortophotos, WorldView-3, and GoogleEarth. Depending on the source, the images have different sizes and pixel Ground Sampling Distances (GSD), as we can see in the Table 1, so we need to standardize to a uniform dimension and GSD.

Source	GSD (cm/pixel)	Image Shape (pixel)
AGEA	20	1050x1050
WorldView3	30	700x700
GoogleEarth	21	1000x1000

Table 1: Different sources with their GSDs and dimensions.

Lastly, in the testing set, there are 169 images accompanied by segmentation annotations. Given that we resort to a weakly supervised approach, these annotations are used only to evaluate the localization results.

3. Proposed solution

3.1. Standardizing Image Dimensions and Ground Sampling Distance

To ensure effective model training, it was essential for all images to have the same dimensions (pixels) and GSD (cm/pixel), so that the wastes depicted in the images are at a consistent scale, allowing us to leverage the maximum GSD possible. Consequently, models are trained more effectively and produce better inferences.

To achieve this, we utilize our knowledge that each image covered an area of 210 × 210 square meters. By dividing one side (21,000 centimeters) by the image size in pixels, we could calculate the exact GSD of each image.

Knowing the original GSD of each image and defined the target GSD, we can resize all images to specific dimensions in order to bring all of them to the target GSD. The specific dimensions for each side is computed with the following formula:

$$\text{newShape} = \text{Shape} \times \left(\frac{\frac{21000cm}{\text{Shape}}}{\text{fixedGSD}} \right)$$

However, this would still leave us with images of varying dimensions. To address this, we apply zero padding around each image to ensure uniform dimensions. Table 2 illustrates the final dimensions of the images at specific GSD:

GSD (cm/pixel)	Image Shape (pixel)
50 cm/pixel	448x448
30 cm/pixel	736x736
20 cm/pixel	1088x1088

Table 2: Different GSDs and their respective dimensions with padding.

3.2. Data Augmentation

Augmentation transformations serve distinct purposes in creating new data. Given limited positive labels, augmenting data is crucial to expand the dataset and enhance model performance, aiming to simulate real satellite scenarios. The transformations used are:

- Random Flip: This transformation randomly flips images horizontally or vertically.
- Random Rotation: By rotating images within a specified angle range.
- Random Noise: Adding random noise to images.

- Random Brightness: This transformation randomly adjusts image brightness.
- Random Contrast: By randomly adjusting image contrast.

In addition, we implemented a strategy inspired by [7] to address a common limitation in class activation maps. This involves randomly obscuring patches of images during training to prompt the model to explore various regions for accurate classification. We divided images into grids and obscured patches based on a probability threshold. Patch sizes of 16x16 and 32x32 were chosen, and probability thresholds of 0.9, 0.8, and 0.7 were tested to minimize false positives.

3.3. HeatMap-Based Weakly Localization Approach

Grad-CAM++ As our initial approach, we trained a binary classifier composed of a Convolutional Neural Network (CNN), Global Average Pooling, and several Dense Layers for the classification head. This enabled us to pinpoint the regions of the image that led the network to classify it as positive, thereby detecting the presence of waste. This method is known as Class Activation Maps (CAM), which utilizes trained classification models to subsequently extract relevant areas. Specifically, we focused on Grad-CAM++, an enhancement of CAM that provides increased precision in localizing salient features and frees us from the architectural constraint of a single dense layer imposed by CAM [1]. The output of Grad-CAM++ is a weighted sum of the channels of the last feature map of the CNN where these weights indicate the importance of each channel in the model's output for a specific class of interest.

To accomplish this, we train a binary classifier using ResNet50 as backbone and from the last feature map of the backbone we compute the GRAD-CAM++.

Heatmap Generation from Feature Maps Across Different Scales The second approach we followed is derived from [[5] and consists in extract feature maps at various scale from the ResNet50 backbone. We apply 1x1 convolution to the output of the last four blocks of the backbone, followed by a sigmoid function, as shown in Figure 1, where h2-h5 are the feature maps, s2-s5 are the extracted heatmaps and y2-y5 are their predicted probabilities. Then we derive classification probability from each heatmap. To enhance accuracy, we employ Top-T Pooling, which averages the first T highest probabilities in the heatmaps.

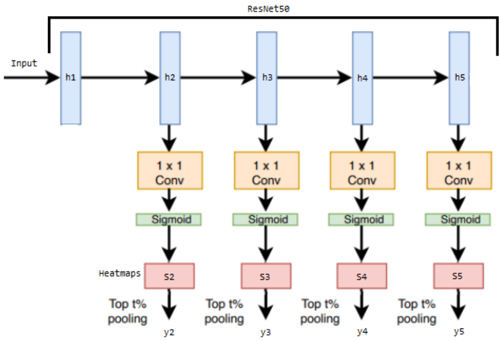

Figure 1: Heatmap Generation from Feature Maps across different Scales.

$$y_n = \frac{1}{|H|} \sum_{(i,j)\in H} S_n(i,j)$$

where:

- y_n the predicted probability extracted from the heatmap n.
- $|H|$ the set containing the locations of the T highest probabilities in the heatmap.
- $S_n(i,j)$ the activation value of pixel (i,j) in the heatmap n.

The training loss function of the model involves computing errors for updating model parameters based on the 4 predictions. Alongside computing cross-entropy error for these predictions, we introduce L1 regularization to enforce sparsity in the heatmaps. This is achieved by summing the absolute values of probabilities from the 4 heatmaps.

$$\sum_{n=2}^{5}(\text{BCE}(y, y_n) + \lambda \sum_{i=1}^{H}\sum_{j=1}^{W}|S_n(i,j)|)$$

where:

- n the index of the heatmap (from 2 to 5).
- BCE the binary cross entropy between the labels y and predictions y_n for heatmap n.
- λ the regularization coefficient for sparsity.
- H and W are the size of the heatmap.
- $S_n(i,j)$ the value of pixel (i,j) in the heatmap n.

After training, when given a high-GSD image, the model generates four heatmaps at different scales with corresponding predictions. The heatmaps are resized to match the original image's GSD, and their average is computed.

3.4. MIL-Based Weakly Localization Approach

Multi-instance learning (MIL) views data instances, like images, as sets of sub-instances rather than individual entities. In our context, the HxWx3 image is the base instance linked to a binary label for waste presence. Extracted patches from this image form sub-instances, collectively constituting a bag associated with the original label.

In our implementation, we considered two scenarios for patch extraction:

- We divided the initial image of size HxWx3 into a grid, where each cell represents a PxPx3 patch. The corresponding bag for an image consists of all non-overlapping extracted patches.
- We selected a subset of the patches using the heatmap obtained from the method in Figure 1 and previously described. Computed the heatmap from the image, we extract patches and associate to each one the corresponding area of the heatmap. A score of the patch is computed as the average over all pixels of the corresponding heatmap area. The patches are ordered based on the computed score in descending order and the top K are selected.

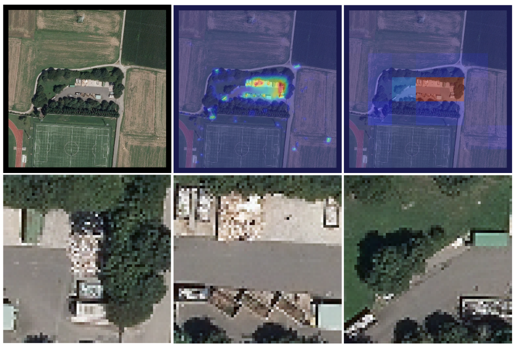

Figure 2: Example of the second scenario.

As shown in Figure 3, each patch P_i in the bag is processed with ResNet50, yielding one-dimensional representations z_k through Global Average Pooling. These representations z_k, K in total, undergo aggregation using a gated attention mechanism from [4]. This mechanism assigns importance to each patch, allowing weighted averaging to form a single vector H. The weights a_k , determined by a neural network, sum to 1 for bag size invariance. The weights for each patch are calculated as follows:

$$a_k = \text{softmax}(w^\top \tanh(V z_k^\top) \odot \sigma(U z_k^\top))$$

where:

- $\odot$ element-wise product.
- a_k are the weights associated with each patch k
- z_k is the embedding (representation) of patch k.
- w, U and V are parameters

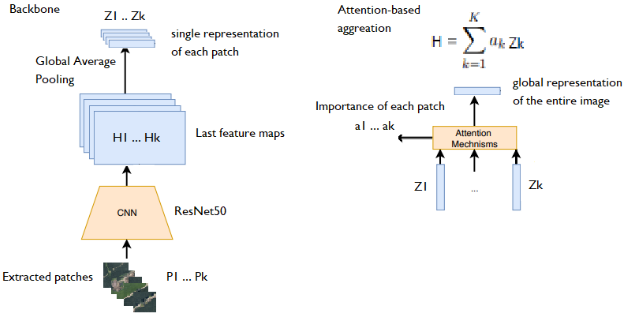

Figure 3: MIL-based framework.

After pooling the K 1D vectors z_k of K patches, a single representative 1D vector H for the patch bag is obtained and fed into the classification head, typically composed of Dense Layers, to predict waste presence or absence probability.

During inference, any number of patches can form the bag due to softmax's bag size independence. Each patch undergoes classification by constructing bags of single patches. A patch is processed through ResNet50, yielding a 1D vector representation. Its weight is calculated and normalized through softmax, preserving the 1D vector for the classification head to provide the patch's probability.

For each patch, its weight before normalization and its probability are determined to create a map of relevant areas in the original image. Each possible patch is extracted, possibly with a reduced shift for precise localizations. Its importance weight and probability of containing waste are computed. Pixels within a patch are then assigned its probability or importance weight, or a fusion of the two. In overlapping patches, pixel values are summed and averaged across patches for an average score

4. Experiments and Results

In this research, we evaluated waste localization performance alongside binary classification indicating waste presence or absence in images. The main metric we considered for classification is F1-Score, while for localization is the Intersection over Union (IoU).

We initially standardized images to a single dimension and GSD by training a CNN followed by Global Average Pooling and Dense Layers, with the last dense layer utilizing the sigmoid function. Additionally, we utilized the Vision Transformer Base with 16x16 pixel patches. We tested ResNet50 and InceptionResNetV2 backbones with images of GSDs 20 cm/pixel, 30 cm/pixel, and 50 cm/pixel (limited to 30 cm/pixel for ViTB16 due to computational constraints).

Architecture	GSD	F1
ResNet50	50 cm/pixel	82.66%
Inception ResNet V2	20 cm/pixel	86.20%
ViT-B16	30 cm/pixel	89.07%

Table 3: Different architecture with their best performances and GSDs.

Results (Table 3) showed InceptionResNetV2 and ViTB16 achieving best performance at maximum achievable GSDs, i.e., 20 cm/pixel and 30 cm/pixel respectively, while ResNet50 was limited to 50 cm/pixel, indicating a limitation for ResNet50 in fully utilizing 20 cm/pixel GSD.

Despite ResNet50's limitation, we focused on it for comparison with a previous study conducted within the PERIVALLON project [8] and computational reasons. We improved classification and localization by combining heatmaps at different scales from different backbone layers. This allowed better object boundary identification with early layers and area localization with last layers, enhancing task performance. Early layer predictions also facilitated relevant feature learning from outset, improving classification. Below, we report the best results obtained with the proposed approach with and without Patch Obscuring, compared to GRAD-CAM++ applied to the previous architecture composed by ResNet50 as backbone, GAP, and dense layers.

Method	**50 cm/p. res.**		**20 cm/p. res.**	
	F1	**IoU**	**F1**	**IoU**
Grad-CAM++	82.66%	13.87%	82.45%	18.87%
Heatmap Generation	84.98%	21.72%	81.86%	24.15%
Heatmap Generation with Patch Obscuring	85.53%	21.94%	82.10%	24.22%

Table 4: Different methods with their best performances at different GSD.

Finally, we evaluated a Multiple Instance Learning (MIL) approach. Instead of predicting waste presence in a single image, we assess it in a set of patches extracted from the entire original image. Each patch undergoes individual processing via ResNet50, resulting in 1D vectors representing their local features. Subsequently, we establish a global context by computing a weighted average of the individual patch representations, where the weights are determined by the importance of each patch calculated using the gate attention mechanism. Finally, the global representation is input to dense layers for the classification of the entire image. Concerning the segmentation task, during inference, we can flexibly utilize any number of patches to construct bags due to the bag size independence of softmax. Thus, each patch can be classified by forming bags comprising single patches. This methodology enables us to construct a heatmap by assigning the probability associated with each pixel in a specific patch. In overlapping patches, pixel values are aggregated and averaged across patches to obtain an overall score. This method enables inference on small patches (e.g., 32x32 or 64x64 pixels). Results showed that patch set cardinality significantly impacts performance; too small a K value (the number of patches in the bag) may miss relevant information, while a too large one risks confusing the model. Larger patch sizes improve classification as set cardinality decreases, aiding model training. However, patches larger than 32x32 aren't beneficial for waste object localization due to their size. In the Table 5, we report the best results obtained with patch sizes of 32 and 64, along with their respective best F1 and IoU scores and the cardinality of the set of the patches K.

Patch Size	50 cm/p. res.			20 cm/p. res.		
	K	F1	IoU	K	F1	IoU
32x32	57	78.48%	20.84%	346	76.75%	26.45%
64x64	15	80.71%	19.44%	86	79.24%	24.79%

Table 5: MIL-based approach with different patch size and GSD.

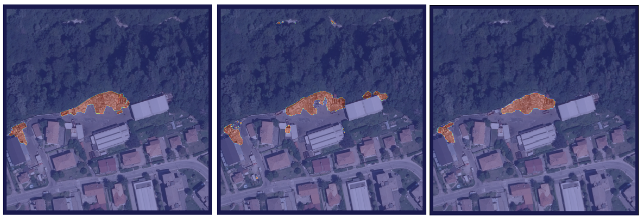

Figure 4: The first mask is the annotated mask while the following masks are those generated with the heatmap-based and MIL-based approaches, respectively

5. Conclusions

The research emphasizes the importance of optimal pre-processing for satellite images, favoring maximum GSD for enhanced performance, though not all architectures may excel under these conditions. High GSDs demand complex architectures and computational resources, while medium-sized models offer efficiency but with spatial limitations. Hence arises the need to investigate the feasibility of using smaller images with reduced context and maximum GSD. In addressing waste segmentation challenges, particularly due to low-GSD feature maps and a focus on discriminative regions for class prediction, the heatmap-based approach encounters obstacles. Nevertheless, DeepLab and similar architectures mitigate these issues by regulating receptive fields without escalating parameters. Additionally, various literature papers delve into mining/iterative methodologies to extract all pertinent elements in an image.

In contrast, while the MIL-based approach detects waste boundaries using small patches, it faces challenges in achieving high precision. Addressing this necessitates considering patch positions and importance weights by incorporating information from other patches rather than processing them independently, as explored in the literature on Transformers [9].

References

[1] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, March 2018.

[2] European Commission CORDIS. PERIVALLON Protecting the EuRopean terrItory from organised enVironmentAl crime through inteLLigent threat detectiON tools. https://cordis.europa.eu/project/id/101073952, 2022-2025.

[3] Piero Fraternali, Luca Morandini, and Sergio Luis Herrera González. Solid waste detection in remote sensing images: A survey, 2024.

[4] Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning, 2018.

[5] Kangning Liu, Yiqiu Shen, Nan Wu, Jakub Chłędowski, Carlos Fernandez-Granda, and Krzysztof J. Geras. Weakly-supervised high-resolution segmentation of mammography images for breast cancer diagnosis, 2021.

[6] PERIVALLON. Introduction to PERIVALLON Project. https://perivallon-he.eu/introduction-to-perivallon-project-2/, 2022-2025.

[7] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization, 2017.

[8] Rocio Nahime Torres and Piero Fraternali. Aerialwaste dataset for landfill discovery in aerial and satellite images. *Scientific Data*, 10(1):63, Jan 2023.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.